A Multi-task Deep Network for Person Re-identification

Chen, Weihua; Chen, Xiaotang; Zhang, Jianguo; Huang, Kaiqi

Published in:
Proceedings of The Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Publication date:
2017

Document Version
Peer reviewed version

To appear in Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), San Francisco, USA, 2017

A Multi-task Deep Network for Person Re-identification

Weihua Chen[1], **Xiaotang Chen**[1], **Jianguo Zhang**[3], **Kaiqi Huang**[1,2,4]
[1]CRIPAC&NLPR, CASIA [2]University of Chinese Academy of Sciences
[3]Computing, School of Science and Engineering, University of Dundee, United Kingdom
[4]CAS Center for Excellence in Brain Science and Intelligence Technology
Email:{weihua.chen, xtchen, kqhuang}@nlpr.ia.ac.cn, j.n.zhang@dundee.ac.uk

Abstract

Person re-identification (ReID) focuses on identifying people across different scenes in video surveillance, which is usually formulated as a binary classification task or a ranking task in current person ReID approaches. In this paper, we take both tasks into account and propose a multi-task deep network (MTDnet) that makes use of their own advantages and jointly optimize the two tasks simultaneously for person ReID. To the best of our knowledge, we are the first to integrate both tasks in one network to solve the person ReID. We show that our proposed architecture significantly boosts the performance. Furthermore, deep architecture in general requires a sufficient dataset for training, which is usually not met in person ReID. To cope with this situation, we further extend the MTDnet and propose a cross-domain architecture that is capable of using an auxiliary set to assist training on small target sets. In the experiments, our approach outperforms most of existing person ReID algorithms on representative datasets including CUHK03, CUHK01, VIPeR, iLIDS and PRID2011, which clearly demonstrates the effectiveness of the proposed approach.

Introduction

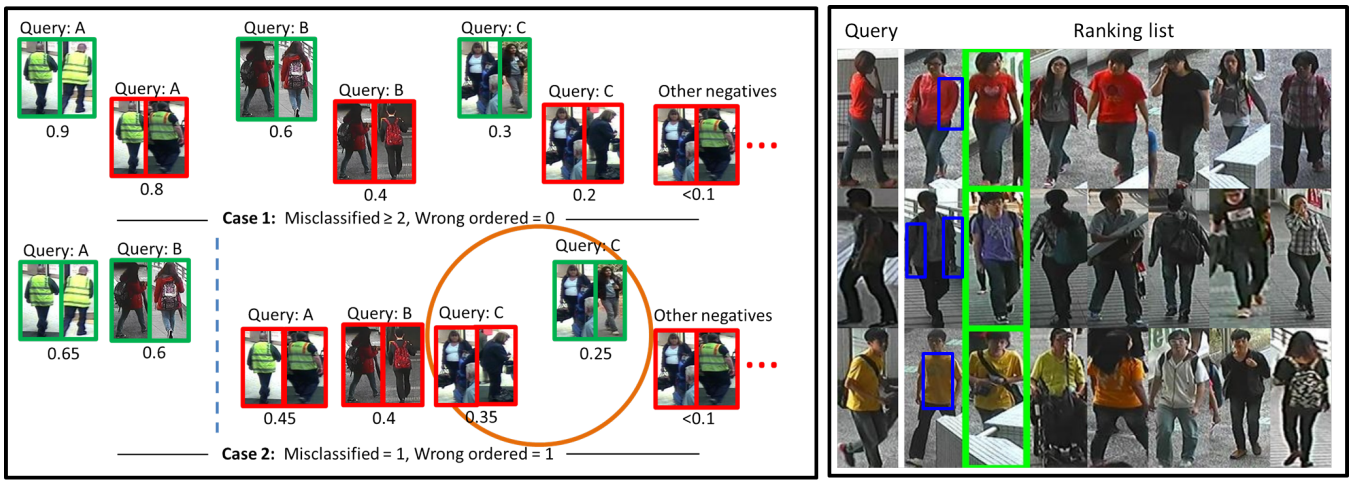
Person re-identification (ReID) is an important task in wide area video surveillance. The key challenge is the large appearance variations, usually caused by the significant changes in human body poses, illumination and camera views. It has many applications, such as inter-camera pedestrian tracking and human retrieval.

Recently, deep learning approaches (Li et al. 2014; Ahmed, Jones, and Marks 2015; Wang et al. 2016) are successfully employed in person ReID with significant performance, especially on large datasets, such as CUHK03. Most deep learning methods (Li et al. 2014; Yi, Lei, and Li 2014; Ahmed, Jones, and Marks 2015) solve the problem as a binary classification issue and adopt a classification loss (*e. g.* a softmax loss) to train their models. The core behind these approaches is to learn identifiable features for each pair for classification. The binary classification loss is usually designed to require all positive pairs should hold smaller distances than all negative pairs. However, in person ReID, we don't have to require all positive pairs holding smaller distances than all negative pairs *regardless of query images*. Instead, what we want is *for each query image*, its positive pairs have smaller distances than its negative ones. Therefore, in some cases[1], the application of binary classification loss may lead the learned model to an undesired locally optimal solution, which is elaborated as below.

The example is shown in Fig. 1 (a). Case 1 and 2 illustrate two projected distributions of scores obtained by trained binary classifiers. For each pair sample, the score underneath denotes the similarity probability between its two images. Query:*X* indicates where an image from person *X* is used as a query image (the left image in a pair). For example, Query:A means an image from person A is used as a query image. Green-coloured rectangle indicates a positive pair, and red rectangle for the negative pair. In Case 1, it is evident that for each query image (w.r.t one particular person), we can get the correct rank-1 match, *i. e.* two images within its positive pairs always hold larger similarity score than those within its negative pairs. However, in this case it is very difficult for a classifier to determine a suitable threshold to get a low misclassification cost (*e. g.* less than two misclassified samples). On the contrary in Case 2, where the vertical dashed line denotes the decision threshold learned by the classifier, the classifier has a lower misclassification rate. As a result, a binary classifier will favor Case 2 rather than Case 1, as the classification loss in Case 2 will be lower than that in Case 1. But in ReID, we prefer Case 1, which outputs correct ranking results for all of the three persons, rather than Case 2 that contains a false rank-1 result (highlighted in an orange circle). Case 2 could be potentially rectified by a ranking loss.

As person ReID commonly uses the Cumulative Matching Characteristic (CMC) curve for performance evaluation which follows rank-*n* criteria, some deep learning approaches (Ding et al. 2015; Chen, Guo, and Lai 2016; Cheng et al. 2016) begin to treat the person ReID as a ranking task, similar to image retrieval, and apply a ranking loss (*e. g.* a triplet loss) to address the problem. The main purpose is to keep the positive pairs maintaining shorter relative distances in the projected space. However, the person ReID differs from image retrieval in that person ReID needs to identify the same person across different scenes (*i. e.* , a

[1]This situation commonly happens when a fixed embedding metric, *e. g.* Euclidean distance, is used for similarity measurement. In this case, it's hard for the network to learn a suitable feature representation.



(a) Classification issue

(b) Rank issue

Figure 1: Problems in two tasks.(a) Classification issue: the classification loss prefer to train a lower misclassification rate model like Case 2 rather than Case 1. (b) Ranking issue: the appearance of top-rank images is more similar to the query image, while the true positive presents a much less similar appearance. (Best viewed in color and see main text for detailed explanation)

task of predicting positive and negative pairs, focusing on identifiable feature learning, and a positive pair is not necessarily the most similar pair in appearance). Ranking-based approaches are sensitive to their similarity measurements. The current measurements (*e. g.* the Euclidean distance in the triplet loss) care more about the similarity to query images in appearance. In the projection space obtained by a model trained on the triplet loss, it's very challenging to find out a true positive which holds a less similar appearance. As shown in Fig. 1 (b), there are three query images. Each has a ranking list returned by a ranking loss, and the left-most is the most similar one to the query. The green rectangle indicates the positive pair (*ground truth*). We can observe that the image ranked first w.r.t each query image is a mismatched image but holding a more similar appearance to the query image than the matched does.

In the person ReID, either the binary classification loss or the ranking loss has its own strengths and weaknesses. As two tasks handle the person ReID from different aspects, we take both of them into account and build a more comprehensive person ReID algorithm. In our method, two tasks are jointly optimized in one deep network simultaneously. We set the binary classification loss and the ranking loss on different layers according their own advantages. The ranking loss encourages a relative distance constraint, while the classification loss seeks to learn discriminative features for each pair during the similarity measurement. As the classification task focuses on feature of pairs, we import the joint feature maps to represent the relationships of paired person images.

Meanwhile, deep learning approaches, such as convolutional neural networks (CNN), benefit a lot from a large scale dataset (*e. g.* ImageNet). However, this is not the case in person ReID. Since manually labeling image pairs is tedious and time-consuming, most of current ReID datasets are often of limited sizes, *e. g.* CUHK01 (Li, Zhao, and Wang 2012), VIPeR (Gray, Brennan, and Tao 2007), iLIDS (Zheng, Gong, and Xiang 2009) and PRID2011 (Hirzer et al. 2011). It could hinder the attempts to maximize the learning potential of our proposed network on each of those datasets. This case can be migrated by using some auxiliary datasets. However, the variations across camera views are different from dataset to dataset. As a consequence, the data of the auxiliary dataset can't be directly used to train models on small datasets. In this paper, the problem is considered as a semi-supervised cross-domain issue (Ganin and Lempitsky 2015). The target domain is the small dataset that contains only a few samples and the source domain is an auxiliary dataset which is large enough for training CNN models. As person ReID can be considered as a binary classification problem, our purpose is to keep the samples of the same class in different domains closer. A cross-domain architecture is further proposed to minimize the difference of the joint feature maps in two datasets, which are belonged to the same class of pairs (*i. e.* , positive pair and negative pair), and utilize the joint feature maps of the auxiliary dataset to fine tune those of small datasets during the training process. In this case, the joint feature maps of small datasets are improved with the data of the auxiliary dataset and boost the ReID performance on smaller target datasets.

In summary, our contributions are three-fold: 1) a novel multi-task deep network for person ReID, where two tasks focuses on different layers and are jointly optimized simultaneously for person ReID; 2) a cross-domain architecture based on the joint feature maps to handle the challenge of limited training set; 3) a comprehensive evaluation of our methods on five datasets, and showing the superior perfor-

mance over most of state-of-the-art methods.

Related work

Most of existing methods in person ReID focus on either feature extraction (Zhao, Ouyang, and Wang 2014; Su et al. 2015; Matsukawa et al. 2016), or similarity measurement (Li and Wang 2013; Shen et al. 2015; Liao and Li 2015). Person image descriptors commonly used include color histogram (Koestinger et al. 2012; Li and Wang 2013; Xiong et al. 2014), local binary patterns (Koestinger et al. 2012), Gabor features (Li and Wang 2013), and etc., which show certain robustness to the variations of poses, illumination and viewpoints. For similarity measurement, many metric learning approaches are proposed to learn a suitable metric, such as locally adaptive decision functions (Li et al. 2013), local fisher discriminant analysis (Pedagadi et al. 2013), cross-view quadratic discriminant analysis (Liao et al. 2015), and etc. A few of them (Xiong et al. 2014; Paisitkriangkrai, Shen, and Hengel 2015) learn a combination of multiple metrics. However, manually crafting features and metrics require empirical knowledge, and are usually not optimal to cope with large intra-person variations.

Since feature extraction and similarity measurement are independent, the performance of the whole system is often suboptimal compared with an end-to-end system using CNN that can be globally optimized via back-propagation. With the development of deep learning and increasing availability of datasets, the handcrafted features and metrics struggle to keep top performance widely, especially on large scale datasets. Alternatively, deep learning is attempted for person ReID to automatically learn features and metrics (Li et al. 2014; Ahmed, Jones, and Marks 2015; Wang et al. 2016). Some of them (Ding et al. 2015; Chen, Guo, and Lai 2016; Cheng et al. 2016) consider person ReID as a ranking issue. For example, Ding *et al.* (Ding et al. 2015) use a triplet loss to get the relative distance between images. Chen *et al.* (Chen, Guo, and Lai 2016) design a ranking loss which minimizes the cost corresponding to the sum of the gallery ranking disorders. Cheng *et al.* (Cheng et al. 2016) add a new term to the original triplet loss function to further constrain the distances of pairs.

Other approaches (Li et al. 2014; Ahmed, Jones, and Marks 2015; Wu et al. 2016) tackle the person ReID problem from the classification aspect. For instance, Yi *et al.* (Yi, Lei, and Li 2014) utilize a siamese convolutional neural network to train a feature representation. Li *et al.* (Li et al. 2014) design a deep filter pairing neural network to solve the ReID problem. Ahmed *et al.* (Ahmed, Jones, and Marks 2015) employ a local neighborhood difference to deal with this misalignment issue. All of them employ a binary classification loss to train their models. It is worth mentioning that there are some papers (Wu et al. 2016; Xiao et al. 2016) using multi-class classification instead of binary classification. They classify identities to solve the person ReID problem, which shares a similar idea with DeepID in face recognition (Sun et al. 2014). However, in most person ReID datasets, there are few samples for each identity. VIPeR (Gray, Brennan, and Tao 2007) and

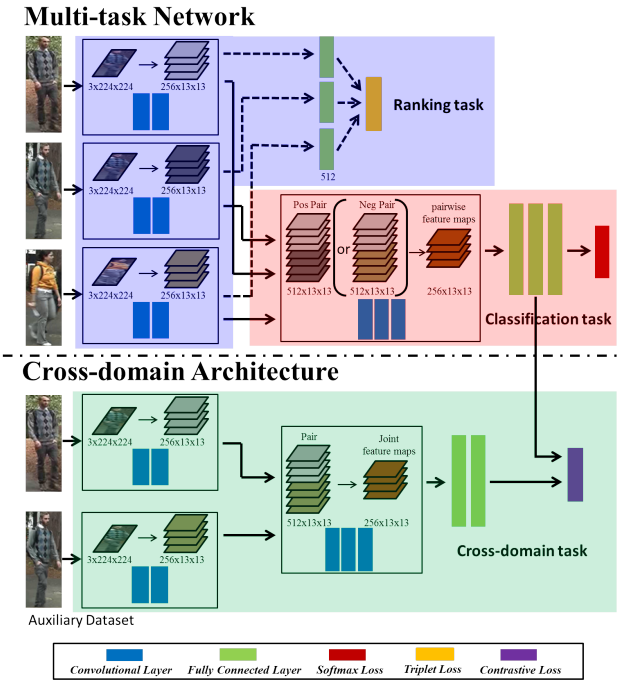


Figure 2: The framework of the proposed multi-task deep network and the cross-domain architecture. The cross-domain architecture is only used when an auxiliary dataset is needed for training.

PRID2011 (Hirzer et al. 2011) datasets have only two images for each person. The lack of training samples may make the multi-class classification less effective. Xiao *et al.* (Xiao et al. 2016) achieve a good performance, but it combines all current datasets together as its training data.

Our network considers two tasks (the classification loss and the ranking loss) simultaneously and takes both of their advantages during training. Wang *et al.* (Wang et al. 2016) also discuss both classification and ranking losses, however, it trains two losses separately and combines them on the score level. In this paper, we jointly optimize two tasks simultaneously in our network.

It is worth noting that none of the works above in person ReID seeks to solve the problem of "learning a deep net on a small dataset" which is a typical case in person ReID. This paper addresses this issue by proposing a cross-domain deep architecture capable of learning across ReID datasets.

The proposed network

The multi-task network

In our method, we build our architecture according to the different focuses of two tasks. As we known, the ranking task concentrates on the orders of images with the same query. Its purpose is to rank the similarities of images and obtain a good ranking list for each query. For two person images, in order to compute their similarity score, we have to compare each part of two people. We can't obtain their similarity score only based on some local parts. In other words,

the global features of the whole images should be paid more attention than local parts during ranking (Tolias, Sicre, and Jegou 2016). Meanwhile, in the association, the most important purpose of the classification task is to distinguish two categories and make the learned features more identifiable. As shown in Fig. 1 (b), the possible key to distinguish the top 1 result from the query is mainly on the blue local regions, *e. g.* using the feature of the sleeves or the belting. So the classification loss should pay more attention on learning these local semantic features, which hold enough identifiable information. In this way, the classification loss would prefer to semantic local features instead of the global features during training.

From Wang's work (Wang et al. 2015), it had been shown that the higher layers in deep network capture semantic concepts, whereas lower layers encode features to capture intra-class variations. For ranking, we compare images based on a combination (global appearance oriented) of low-level features (*i. e.* edges, bars etc) learned in lower layers to overcome intra-class variations (as suggested by Wang's work (Wang et al. 2015)). Features in high layers focus on identifiable local semantic concepts, driven by the classification loss. The whole framework is shown in Fig. 2. The ranking loss provides global low-level features which could be appropriate for image similarity ranking, and the classification loss further learns the identifiable local features based on the low-level ones. Then we give the details of our multi-task network.

The ranking part is a triplet-input model. For each positive pair, we produce ten triplets (a positive pair + a negative image: A_1, A_2, B_2 [2]). All these triplets constitute our training data. The input triplet contains three images, each of the size $3 * 224 * 224$. The ranking task includes two convolutional layers at the beginning, which are used to reinforce the learning of global features. After the two convolutional layers, three sets of feature maps hold the same size of $256 * 13 * 13$ and are sent to a triplet loss through a shared fully connected layer. The triplet loss being minimized is the same as FaceNet (Schroff, Kalenichenko, and Philbin 2015):

$$L_{trp} = \sum_{i=1}^{N} [\|f_{A1} - f_{A2}\|_2^2 - \|f_{A1} - f_{B2}\|_2^2 + \alpha]_+ \quad (1)$$

where α is a margin that is enforced between positive and negative pairs, N is the number of the triplets. $f \in \mathbb{R}^{512}$ denotes the features input to the triplet loss from three images. Minimizing the triplet loss is to reserve the information of relative distances between input images.

In the classification part, the input of the third convolutional layer is a set of feature maps of an image pair. The three sets of feature maps with the size of $256 * 13 * 13$ from the ranking task are regrouped into two types of pairs, a positive pair and a negative pair. The feature maps from the two images of the same person, *i. e.* (A_1, A_2), are concatenated as a positive pair, while one image in the positive pair (A_1) and one negative image (B_2) from the different camera view are stacked to form the negative pair. The size of feature maps of each pair is $512 * 13 * 13$. These two pairs are fed to three convolutional layers in order, one at each time. The feature maps learned from these layers are called the joint feature maps, which come from each input pair to encode the relationship of two images. Then they are sent into the full connected layers to calculate the similarity. The joint feature maps hold the identifiable information of the input image pair that can represent the relationship of two images. We use these joint feature maps to identify whether the input image pair is from the same person. The classification loss in our network is the binary logistic regression loss, the same as the binary softmax loss in (Li et al. 2014; Ahmed, Jones, and Marks 2015):

$$L_{cls} = -\sum_{i=1}^{N} [(1-y)p(y=0|x) + yp(y=1|x)] \quad (2)$$

where $y \in \{0, 1\}$. When the input pair is a positive pair (*e. g.* (A_1, A_2)), $y = 1$. On the contrary, $y = 0$ for a negative pair (*e. g.* (A_1, B_2)). $p(y|x)$ is the discrete probability distribution over two categories $y \in \{0, 1\}$.

Our five convolutional layers are extended from the architecture of AlexNet (Krizhevsky, Sutskever, and Hinton 2012), differing in that the size of each kernel in the third convolutional layer is $(512 \times 3 \times 3)$ instead of $(256 \times 3 \times 3)$ used in AlexNet. In the train phase, the triplet loss optimises the first two convolutional layers while the classification loss simultaneously trained all five convolutional layers including the first two. In other words, the kernels of the first two layers are jointly optimised by two losses for extracting a global feature of each image. The left three layers are mainly trained by the classification loss to obtain an identifiable feature for image pairs to achieve the binary person identification. In the test phase, only the classification task architecture (including the first two layers) is used. The input two images are sent through five convolutional layers and three fully connected layers, with the last layer predicting the similarity probability of a test pair.

Cross-domain architecture

For most person ReID datasets, the size of data is too small to train a deep model. The common way is to crop or mirror the images, which can increase the number of samples in datasets. However, even with these augmentation processes, the total number of the samples is still far from the requirement of deep learning. This problem is considered as a semi-supervised cross-domain issue in this paper. In cross-domain transfer, the assumption is that two domains share the same task but the data distributions are different. For example, in image classification, two domains would have the same category but the images contain different views or illuminations. In our issue, the corresponding assumption is that two ReID datasets should share the same similarity function while different variations caused by views or poses widely exist in images from two datasets.

In Fig.2, the relationship of two images is reflected by the joint feature maps. For two positive pairs from two different datasets, the learned similarity metrics for each of the pairs

[2] A, B are the person IDs and $1, 2$ mean the camera IDs.

should ideally lead to the same prediction results, *i. e.* both of the pairs are matched pairs. To achieve such a transfer, we propose to force the learned joint feature maps of positive pairs from two datasets closer than those of negative pairs.

The proposed cross-domain architecture is also shown in Fig.2, which utilizes a contrastive loss (Chopra, Hadsell, and LeCun 2005) to keep the two sets of joint feature maps of the same class as similar as possible during the training process. The label for the two pairs is designed as following:

$$label_p = label_a \odot label_b \tag{3}$$

where $\odot$ means the XNOR operation, $label_a \in \{0, 1\}$ is the label for a pair from source; $label_b \in \{0, 1\}$ is the label for a pair from target; $label_p$ is the result after performing the XNOR operation between the labels of those two pairs. If the labels of the two pairs are the same (*i. e.* $label_a$ and $label_b$ are the same), the contrastive loss will keep the two sets of the joint feature maps closer, and otherwise farther. The loss is as following:

$$\begin{gathered} L_{cts} = -\sum_{i=1}^{N} [y\frac{1}{2}d_w^2 + (1-y)\frac{1}{2}max(0, m - d_w)^2] \\ d_w = \|F_a - F_b\|_2 \end{gathered} \tag{4}$$

where y is the label of two pairs after the XNOR operation, F_a and F_b are responses of the feature maps after the second fully connected layer from two datasets.

The training phase of the cross-domain architecture is also a multi-task process. The softmax loss and the triplet loss are to do the re-identification task, while the contrastive loss is employed to keep two sets of joint feature maps from the same class in two datasets as similar as possible. After training, only the model on the target dataset will be reserved for testing. The whole process can be considered as another kind of fine-tune operation using a cross-domain architecture. The purpose is to use the joint feature maps learned on the auxiliary source dataset to fine tune those on smaller target sets during training and boost the ReID performances.

It is worth noting that we don't force the feature maps of two completely different people, each from one of two datasets, to be similar. Instead we ensure that the way in which image pairs are compared (encoded by the learned weights on the joint feature maps) is similar and could be shared across the two datasets. That is the motivation of importing the cross-domain architecture.

Experiments

We conducts two sets of experiments: 1) to evaluate the proposed multi-task deep net (including single-task nets) and the cross-domain architecture; 2) to compare the proposed approach with state of the arts.

Setup

Implementation and protocol. Our method is implemented using the Caffe framework (Jia et al. 2014). All images are resized to 224×224 before being fed to network. The learning rate is set to 10^{-3} consistently across all experiments. For all the datasets, we horizontally mirror each image and increase the dataset sizes fourfold. We use a pre-trained AlexNet model (trained on Imagenet dataset (Krizhevsky, Sutskever, and Hinton 2012)) to initialize the kernel weights of the first two convolutional layers. Cumulative Matching Characteristics (CMC) curves are employed to measure the ReID performance. We report the single-shot results on all the datasets.

Dataset and settings. The experiment is conducted on one large dataset and four small datasets. The large dataset is CUHK03 (Li et al. 2014), containing 13164 images from 1360 persons. We randomly select 1160 persons for training, 100 persons for validation and 100 persons for testing, following exactly the same setting as (Li et al. 2014) and (Ahmed, Jones, and Marks 2015). The four small datasets are CUHK01 (Li, Zhao, and Wang 2012), VIPeR (Gray, Brennan, and Tao 2007), iLIDS (Zheng, Gong, and Xiang 2009) and PRID2011 (Hirzer et al. 2011). In CUHK01 dataset, we randomly choose only 100 persons for testing, and all the rest 871 persons are used for training. For three other datasets, we randomly divide the individuals into two equal parts, with one used for training and the other for testing. Specifically, in the PRID2011 dataset, besides 100 test individuals, there are another 549 people in the gallery.

Results for the multi-task network

Multi vs. single task. Results of CMCs with different rank accuracies are shown in Table. 1. The proposed multi-task network (Fig. 2) is denoted by *MTDnet*. As *MTDnet* adopts the classification loss for testing, we give results using the ranking loss for testing with the same model (denoted by *MTDtrp*). It's obvious that the performance of *MTDnet* is much better than *MTDtrp* which implies the last three convolutional layers trained with the classification loss indeed provide a great help to increase the person ReID performance. The results of the single-task networks using the triplet ranking loss (denoted by *MTDnet-rnk*) and the binary classification loss (denoted by *MTDnet-cls*) individually are also provided. It is worth noting that, for a fair comparison, the architecture of *MTDnet-rnk* network is expanded into containing five convolutional layers plus three fully connected layers as AlexNet (Krizhevsky, Sutskever, and Hinton 2012) instead of the two convolutional layers shown in Fig. 2, *i. e.* the number of layers in two single-task networks is the same. The similarity of two images in *MTDnet-rnk* is computed with the Euclidean distance. On CUHK03, our multi-task network (*MTDnet*) achieves a rank-1 accuracy of 74.68% and is much better than either *MTDnet-cls* or *MTDnet-rnk*, which indicates the complementarity of two tasks and the effectiveness of jointly optimizing. On four small datasets, our multi-task network consistently outperforms each of two single-task nets (*MTDnet-cls* and *MTDnet-rnk*).

Cross-domain architecture. We compare the cross-domain architecture (*MTDnet-cross*) with the original multi-task network (*MTDnet*) on four small datasets. In this experiment, CUHK03 is considered as the dataset from the source domain, while each of the four small dataset is from the target domain. Therefore, the knowledge transfer is from

Table 1: The CMC performance of the state-of-the-art methods and different architectures in our method on five representative datasets. The bold indicates the best performance.

Method	Type	CUHK03			CUHK01			VIPeR			iLIDS			PRID2011		
		$r=1$	$r=5$	$r=10$	$r=1$	$r=5$	$r=10$	$r=1$	$r=5$	$r=10$	$r=1$	$r=5$	$r=10$	$r=1$	$r=5$	$r=10$
PRDC (Zheng, Gong, and Xiang 2011)	-	-	-	-	-	-	-	15.70	38.40	53.90	37.80	63.70	75.10	4.50	12.60	19.70
SDALF (Farenzena et al. 2010)	-	5.60	23.45	36.09	9.90	41.21	56.00	19.87	38.89	49.37	-	-	-	-	-	-
ITML (Davis et al. 2007)	-	5.53	18.89	29.96	17.10	42.31	55.07	-	-	-	29.00	54.00	70.50	12.00	-	36.00
eSDC (Zhao, Ouyang, and Wang 2013)	-	8.76	24.07	38.28	22.84	43.89	57.67	26.31	46.61	58.86	-	-	-	-	-	-
KISSME (Koestinger et al. 2012)	-	14.17	48.54	52.57	29.40	57.67	62.43	19.60	48.00	62.20	28.50	55.30	68.70	15.00	-	39.00
FPNN (Li et al. 2014)	Cls	20.65	51.00	67.00	27.87	64.00	77.00	-	-	-	-	-	-	-	-	-
mFilter (Zhao, Ouyang, and Wang 2014)	-	-	-	-	34.30	55.00	65.30	29.11	52.34	65.95	-	-	-	-	-	-
kLFDA (Xiong et al. 2014)	-	48.20	59.34	66.38	42.76	69.01	79.63	32.33	65.78	79.72	39.80	65.30	77.10	22.40	46.60	58.10
DML (Yi, Lei, and Li 2014)	Cls	-	-	-	-	-	-	34.40	62.15	75.89	-	-	-	17.90	37.50	45.90
IDLA (Ahmed, Jones, and Marks 2015)	Cls	54.74	86.50	94.00	65.00	89.50	93.00	34.81	63.32	74.79	-	-	-	-	-	-
SIRCIR (Wang et al. 2016)	Cls/Rnk	52.17	85.00	92.00	72.50	91.00	95.50	35.76	67.00	82.50	-	-	-	-	-	-
DeepRanking (Chen, Guo, and Lai 2016)	Rnk	-	-	-	70.94	92.30	96.90	38.37	69.22	81.33	-	-	-	-	-	-
DeepRDC (Ding et al. 2015)	Rnk	-	-	-	-	-	-	40.50	60.80	70.40	52.10	68.20	78.00	-	-	-
NullReid (Zhang, Xiang, and Gong 2016)	-	58.90	85.60	92.45	64.98	84.96	89.92	42.28	71.46	82.94	-	-	-	29.80	52.90	**66.00**
SiameseLSTM (Varior et al. 2016)	Cls	57.30	80.10	88.30	-	-	-	42.40	68.70	79.40	-	-	-	-	-	-
Ensembles (Paisitkriangkrai, Shen, and Hengel 2015)	-	62.10	89.10	94.30	53.40	76.30	84.40	45.90	**77.50**	**88.90**	50.34	72.00	82.50	17.90	40.00	50.00
GatedSiamese (Varior, Haloi, and Wang 2016)	Cls	68.10	88.10	94.60	-	-	-	37.80	66.90	77.40	-	-	-	-	-	-
ImpTrpLoss (Cheng et al. 2016)	Rnk	-	-	-	53.70	84.30	91.00	**47.80**	74.70	84.80	**60.40**	**82.70**	**90.70**	22.00	-	47.00
MTDnet-rnk	Rnk	60.13	90.51	95.15	63.50	80.00	89.50	28.16	52.22	65.19	41.04	69.94	78.61	22.00	41.00	48.00
MTDnet-cls	Cls	68.35	93.46	97.47	76.50	94.00	97.00	44.30	69.94	81.96	54.34	73.41	86.13	28.00	50.00	60.00
MTDnet-trp	Cls+Rnk	66.03	84.81	89.87	66.00	84.00	91.50	34.81	60.13	72.78	46.82	72.83	81.50	26.00	49.00	57.00
MTDnet	Cls+Rnk	**74.68**	**95.99**	**97.47**	77.50	95.00	97.50	45.89	71.84	83.23	57.80	78.61	87.28	**32.00**	51.00	62.00
MTDnet-aug	Cls+Rnk	-	-	-	75.50	93.50	97.00	43.35	70.25	78.48	54.91	74.57	84.97	27.00	46.00	59.00
MTDnet-cross	Cls+Rnk	-	-	-	**78.50**	**96.50**	**97.50**	47.47	73.10	82.59	58.38	80.35	87.28	31.00	**54.00**	61.00

CUHK03 to each of the four small datasets. The results of *MTDnet* on four small datasets is obtained by fine tuning the CUHK03 trained model on each small dataset. In the cross-domain architecture, both the target domain network and the source domain network are initialized using the model trained on CUHK03. And in test phase, only the target domain network is used to compute results. Relevant preformance are shown in Table.1. It's obvious that almost all results of the cross-domain architecture are better than those of *MTDnet*, which demonstrates the effectiveness of the cross-domain architecture. We also import another network (*MTDnet-aug*) which simply adds the source data into the target dataset directly and combined them as an augmented dataset for the target dataset training. It's clear that the results of our cross-domain architecture are better than those of *MTDnet-aug*. The models trained with the augmented data (*MTDnet-aug*) are even worse compared with *MTDnet*, which suggests that the direct combination of the source and target datasets is not helpful but disruptive for the training in the target dataset.

Comparison with the state of the arts

We compare ours with representative ReID methods including 18 algorithms, whichever have the results reported on at least one of the five datasets. All of the results can be seen from Table. 1. We have marked all the deep learning methods in the *Type* column. All the non-deep learning approaches are listed as "-". *Cls* indicates deep methods based on the classification loss, while *Rnk* are on the ranking loss. SIRCIR method (Wang et al. 2016) offers the results on both the classification loss and the ranking loss. But in its network, the losses are trained separately. Its combination of two losses are only on the score level, while we jointly optimize two losses in one network and train them simultaneously. Most of these deep methods are in the top performance group among all of the methods considered. It is noted that our results are better than most approaches above, which further confirms that jointly optimizing the two losses has a clear advantage over a single loss. Under the rank-1 accuracy, our multi-task network outperforms all existing person ReID algorithms on CUHK03, CUHK01 and PRID2011. ImpTrpLoss (Cheng et al. 2016) provides the best rank-1 performance on VIPeR and iLIDS. We can see our results are comparable with its, and much better on other datasets.

Conclusion

In this paper, a multi-task network has been proposed for person ReID, which integrates the classification and ranking tasks together in one network and takes the advantage of their complementarity. In the case of having small target datasets, a cross-domain architecture has been further introduced to fine tune the joint feature maps and improve the performance. The results of the proposed network have outperformed almost all state-of-the-art methods compared on both large and small datasets.

References

Ahmed, E.; Jones, M.; and Marks, T. K. 2015. An improved deep learning architecture for person re-identification. In *CVPR*.

Chen, S.-Z.; Guo, C.-C.; and Lai, J.-H. 2016. Deep ranking

for person re-identification via joint representation learning. *TIP* 25(5):2353–2367.

Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; and Zheng, N. 2016. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, volume 1, 539–546.

Davis, J. V.; Kulis, B.; Jain, P.; Sra, S.; and Dhillon, I. S. 2007. Information-theoretic metric learning. In *ICML*.

Ding, S.; Lin, L.; Wang, G.; and Chao, H. 2015. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition* 48(10):2993–3003.

Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; and Cristani, M. 2010. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2360–2367.

Ganin, Y., and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*.

Gray, D.; Brennan, S.; and Tao, H. 2007. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3.

Hirzer, M.; Beleznai, C.; Roth, P. M.; and Bischof, H. 2011. Person re-identification by descriptive and discriminative classification. In *Image Analysis*. Springer. 91–102.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACM on Multimedia*, 675–678.

Koestinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P. M.; and Bischof, H. 2012. Large scale metric learning from equivalence constraints. In *CVPR*, 2288–2295.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, 1097–1105.

Li, W., and Wang, X. 2013. Locally aligned feature transforms across views. In *CVPR*, 3594–3601.

Li, Z.; Chang, S.; Liang, F.; Huang, T. S.; Cao, L.; and Smith, J. R. 2013. Learning locally-adaptive decision functions for person verification. In *CVPR*, 3610–3617.

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 152–159.

Li, W.; Zhao, R.; and Wang, X. 2012. Human reidentification with transferred metric learning. In *ACCV*, 31–44.

Liao, S., and Li, S. Z. 2015. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2197–2206.

Matsukawa, T.; Okabe, T.; Suzuki, E.; and Sato, Y. 2016. Hierarchical gaussian descriptor for person re-identification. In *CVPR*.

Paisitkriangkrai, S.; Shen, C.; and Hengel, A. v. d. 2015. Learning to rank in person re-identification with metric ensembles. In *CVPR*.

Pedagadi, S.; Orwell, J.; Velastin, S.; and Boghossian, B. 2013. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 3318–3325.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 815–823.

Shen, Y.; Lin, W.; Yan, J.; Xu, M.; Wu, J.; and Wang, J. 2015. Person re-identification with correspondence structure learning. In *ICCV*.

Su, C.; Yang, F.; Zhang, S.; Tian, Q.; Davis, L. S.; and Gao, W. 2015. Multi-task learning with low rank attribute embedding for person re-identification. In *ICCV*, 3739–3747.

Sun, Y.; Chen, Y.; Wang, X.; and Tang, X. 2014. Deep learning face representation by joint identification-verification. In *NIPS*, 1988–1996.

Tolias, G.; Sicre, R.; and Jegou, H. 2016. Particular object retrieval with integral max-pooling of cnn activations. In *ICLR*.

Varior, R. R.; Shuai, B.; Lu, J.; Xu, D.; and Wang, G. 2016. A siamese long short-term memory architecture for human re-identification. In *CVPR*.

Varior, R. R.; Haloi, M.; and Wang, G. 2016. Gated siamese convolutional neural network architecture for human re-identication. In *CVPR*.

Wang, L.; Ouyang, W.; Wang, X.; and Lu, H. 2015. Visual tracking with fully convolutional networks. In *ICCV*.

Wang, F.; Zuo, W.; Lin, L.; Zhang, D.; and Zhang, L. 2016. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*.

Wu, S.; Chen, Y.-C.; Li, X.; Wu, A.-C.; You, J.-J.; and Zheng, W.-S. 2016. An enhanced deep feature representation for person re-identification. In *WACV*, 1–8.

Xiao, T.; Li, H.; Ouyang, W.; and Wang, X. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*.

Xiong, F.; Gou, M.; Camps, O.; and Sznaier, M. 2014. Person re-identification using kernel-based metric learning methods. In *ECCV*, 1–16.

Yi, D.; Lei, Z.; and Li, S. Z. 2014. Deep metric learning for practical person re-identification. In *ICPR*.

Zhang, L.; Xiang, T.; and Gong, S. 2016. Learning a discriminative null space for person re-identification. In *CVPR*.

Zhao, R.; Ouyang, W.; and Wang, X. 2013. Unsupervised salience learning for person re-identification. In *CVPR*.

Zhao, R.; Ouyang, W.; and Wang, X. 2014. Learning mid-level filters for person re-identification. In *CVPR*, 144–151.

Zheng, W.; Gong, S.; and Xiang, T. 2009. Associating groups of people. In *BMVC*, volume 2, 6.

Zheng, W.; Gong, S.; and Xiang, T. 2011. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 649–656.